# Selection Bias - What You Don't Know Can Hurt Your Bottom Line.

Gaétan Veilleux, Valen Technologies

2011 CAS Ratemaking and Product Management Seminar

March 20-22, 2011

VALEN
technologies

# Antitrust Notice

- The Casualty Actuarial Society is committed to adhering strictly to the letter and spirit of the antitrust laws. Seminars conducted under the auspices of the CAS are designed solely to provide a forum for the expression of various points of view on topics described in the programs or agendas for such meetings.

- Under no circumstances shall CAS seminars be used as a means for competing companies or firms to reach any understanding – expressed or implied – that restricts competition or in any way impairs the ability of members to exercise independent business judgment regarding matters affecting competition.

- It is the responsibility of all seminar participants to be aware of antitrust regulations, to prevent any written or verbal discussions that appear to violate these laws, and to adhere in every respect to the CAS antitrust compliance policy.

"I don't like statistics.  It's like logic, it doesn't make any sense."

# What Is Selection Bias?

- "A type of bias caused by choosing non-random data for statistical analysis. The bias exists due to a flaw in the sample selection process, where a subset of the data is systematically excluded due to a particular attribute. The exclusion of the subset can influence the statistical significance of the test, or produce distorted results." (Investopedia)

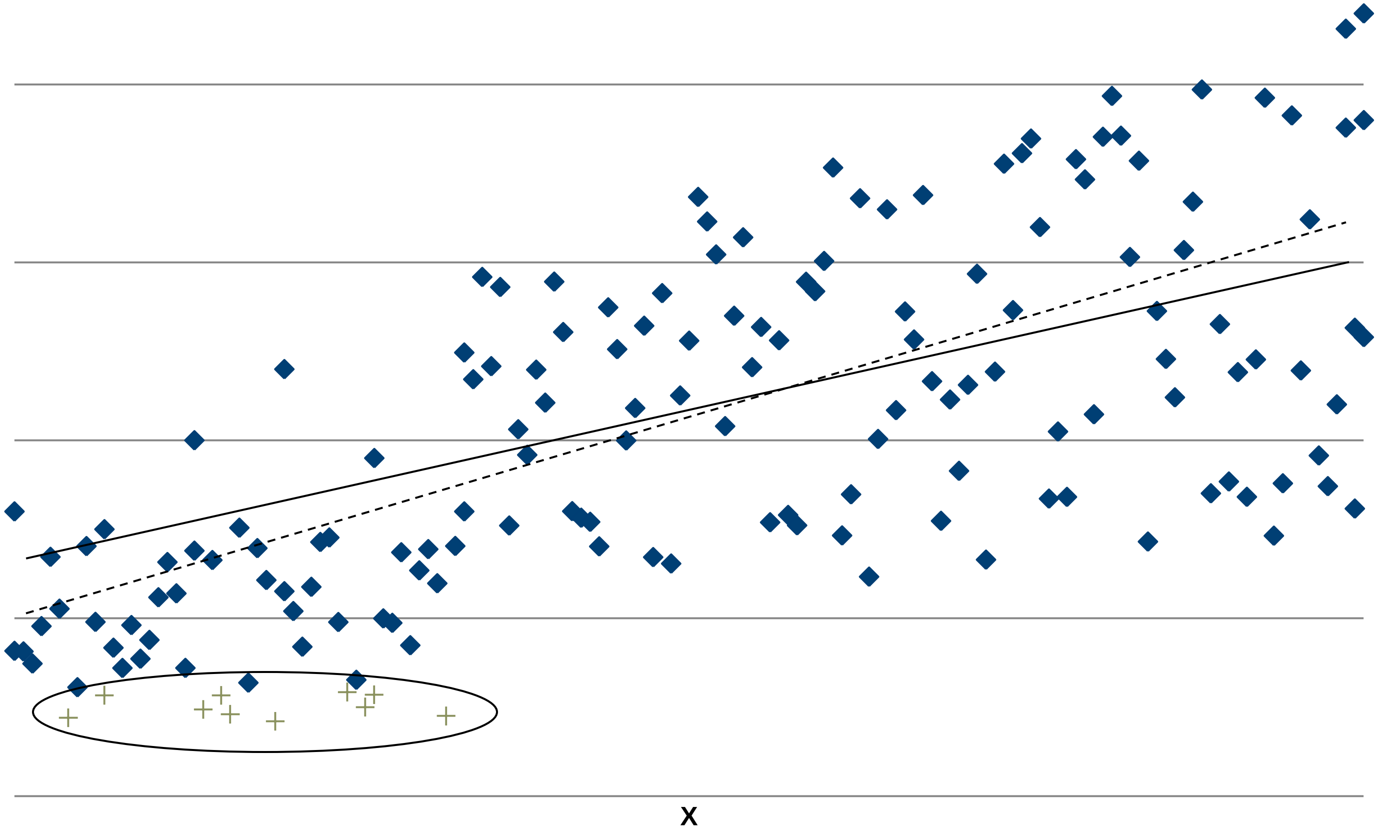
- Selection bias results from estimation on a subsample of individuals who have essentially elected themselves for estimation through their decision to participate in a particular program.
  - Sample selection bias occurs if those who choose not to participate are systematically different from those who do
  - Attrition bias occurs if selected individuals are "lost" over time and those who are lost differ systematically from those who remain.

# Should We Be Concerned?

- The systematic selection of a sub-sample which differs from the overall population will yield distorted empirical results of the population of interest.

- Building a model on such data without attempting to mitigate for the non-random sampling will yield biased estimates or estimates that apply only to the selected sub-sample.

# Example Of Misspecification



# Selection Bias Outside of Insurance

- Economics/Econometrics
- Finance/Credit Industry
- Social Sciences
- Marketing
- Political Science
- Epidemiology
- Investment Analysis
- Insurance
- Many Others

# Selection Bias In Insurance

- Do any insurance processes systematically exclude sub-sets of a population?
  - Pricing
  - Underwriting
  - Claims
  - Marketing
  - Customer service
  - Customer Retention

- What is the source of the systematic selection process?

# Statistical Methods

# 3 Modified Distributional Forms

- Truncation

  A sample is drawn from a subset of a larger population of interest.

- Censoring

  All values above or below some value are set to one value.

- Sample Selection (incidental truncation)

  A specific form of Truncation.

# Truncated Normal Distribution

**Density of a truncated random variable:**

$$f(y|y > a) = \frac{f(y)}{Prob(y > a)} = \frac{\frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right)}{1-\Phi(\alpha)}$$

**Moments:** $E[y|y > a] = \mu + \sigma\lambda(\alpha)$ $\qquad Var[y|y > a] = \sigma^2[1-\delta(\alpha)]$

$$where\ \alpha = (a-\mu)/\sigma$$

$$\lambda(\alpha) = \phi(\alpha)/[1-\Phi(\alpha)\,]$$ ← Inverse Mills Ratio

$$and\ \delta(\alpha) = \lambda(\alpha)[\lambda(\alpha)-\alpha]$$

**Log-Likelihood:** $lnL = \sum_{i=1}^{N}(ln[f(y)] - ln[1-\Phi(\alpha)]$

# Truncated Regression Model

$$y_i = \boldsymbol{x}_{\boldsymbol{i}}'\boldsymbol{\beta} + \varepsilon_i \quad where \quad \varepsilon_i|\boldsymbol{x}_{\boldsymbol{i}} \sim N[0, \sigma^2] \quad and \quad y_i|\boldsymbol{x}_{\boldsymbol{i}} \sim N[\boldsymbol{x}_{\boldsymbol{i}}'\boldsymbol{\beta}, \sigma^2]$$

$$E[y_i|y_i > a] = \boldsymbol{x}_{\boldsymbol{i}}'\boldsymbol{\beta} + \sigma \frac{\phi[(a - \boldsymbol{x}_{\boldsymbol{i}}'\boldsymbol{\beta})/\sigma]}{1 - \Phi[(a - \boldsymbol{x}_{\boldsymbol{i}}'\boldsymbol{\beta})/\sigma]} = \boldsymbol{x}_{\boldsymbol{i}}'\boldsymbol{\beta} + \sigma\lambda(\alpha_i)$$

Marginal effects:

$$\frac{\partial E[y_i|y_i > a]}{\partial \boldsymbol{x}_{\boldsymbol{i}}} = \boldsymbol{\beta} + \sigma(d\lambda_i/d\alpha_i)\frac{\partial \alpha_i}{\partial \boldsymbol{x}_{\boldsymbol{i}}} = \boldsymbol{\beta}\underset{0<\delta<1}{\underline{(1 - \delta_i)}}$$

# Censored data

- Stochastic Censoring - some observations of a dependent variable $y_i$ are censored

  Example 1: The amount a person is willing to spend to buy a car is lower than the least expensive car. There will be no purchase and we do not observe the amount, $y_i$, they would spend.

  Example 2: Losses greater than a loss limit. If a large loss is recorded at the loss limit, the amount above the limit is not available for analysis.

- Tobit model 1958

# Censored Normal Distribution (1)

Define a new y transformed from the latent variable y* as

y = a if y* ≤ a

y= y* if y* > a

**Density of a censored random variable:**

$$f(y) = [f(y^*)]^{d_i}[F(a)]^{1-d_i}$$

**Moments:** $E[y] = \ a\Phi + (1 - \Phi)(\mu + \sigma\lambda)$

$$Var[y] = \sigma^2(1 - \Phi)[1 - \delta) + (\alpha - \lambda)^2\Phi]$$

# Censored Normal Distribution (2)

**Log-Likelihood:**

$$lnL = \sum_{i=1}^{N} \left\{ d_i \left( -ln\sigma + ln\phi \left( \frac{y_i - \mu}{\sigma} \right) \right) + (1 - d_i) ln \left( 1 - \Phi \left( \frac{\mu - a}{\sigma} \right) \right) \right\}$$

**Special Case: a = 0**

$$E[y] = \Phi \left( \frac{\mu}{\sigma} \right) (\mu + \sigma\lambda)$$

where $\lambda = \frac{\phi\left(\frac{\mu}{\sigma}\right)}{\Phi\left(\frac{\mu}{\sigma}\right)}$ (Inverse Mills Ratio)

# Standard Tobit Model

- Assumptions
  - The underlying disturbances are normally distributed
  - The same data generating process that determines the censoring is the same process that determines the outcome variable
  - The dependent variable is censored at zero, i.e.

$$
\begin{aligned}
y &= 0 \quad && \text{if } y^* \leq 0 \\
y &= y^* \quad && \text{if } y^* > 0
\end{aligned}
$$

# Tobit Model

Expected Values of Possible Interest:

1. Expected value of y*, the latent variable

$$E[y^*] = X_i\beta$$

2. E[y|y > 0] – the truncated model

$$E[y|y>0] = \ X_i\beta + \sigma\lambda(\alpha)$$

where $\lambda = \dfrac{\phi\left(\frac{X_i\beta}{\sigma}\right)}{\Phi\left(\frac{X_i\beta}{\sigma}\right)} \quad \mathit{is\ the\ inverse\ Mills\ ratio}$

3. E[y] – the censored model

$$E[y] = \ \Phi\left(\frac{X_i\beta}{\sigma}\right)[X_i\beta + \sigma\lambda(\alpha)]$$

# Tobit Model Estimation

**Log-Likelihood:**

$$lnL = \sum_{i=1}^{N} \left\{ d_i \left( -ln\sigma + ln\phi\left(\frac{y_i - \boldsymbol{y}_i'\boldsymbol{\beta}}{\sigma}\right)\right) + (1-d_i)ln\left(1-\Phi\left(\frac{\boldsymbol{y}_i'\boldsymbol{\beta}}{\sigma}\right)\right)\right\}$$

$$lnL = \sum_{y_i>0} -\frac{1}{2}\left[ln(2\pi) + ln\sigma^2 + \frac{(y_i - \boldsymbol{y}_i'\boldsymbol{\beta})^2}{\sigma^2}\right] + \sum_{y_i=0} ln\left[1-\Phi\left(\frac{\boldsymbol{y}_i'\boldsymbol{\beta}}{\sigma}\right)\right]$$

## Why not use OLS?

- E[y] is non-linear
- OLS estimates of β are inconsistent
- OLS parameters are approximately proportional to Tobit parameters

# Incidental Truncation

- We do not observe y due to the effect of another variable(s)
- A non-random selection process
  - Examples:

    Wage offers are observed only for those who work. Workforce participation may be affected by some unobserved variables which also affect the wage offer.

    Audit results are observed only for audited policies.  The decision to audit specific policies is influenced by other variables, some observed, some not which can affect the audit results.

# Incidental Truncation Distribution

Random variables y and z have a bivariate distribution with correlation ρ.

**Incidentally Truncated joint density of y and z:**

$$f(y, z|z > a) = \frac{f(y, z)}{Prob(z > a)}$$

**Moments of the Incidentally Truncated Bivariate Normal distribution:**

$$E[y|z > a] = \mu_y + \underline{\rho}\sigma_y \lambda(\alpha_z) \qquad Var[y|z > a] = \sigma_y^2[1 - \underline{\rho}^2 \delta(\alpha_z)]$$

$$where\ \alpha_z = (a - \mu_z)/\sigma_z$$

$$\lambda(\alpha_z) = \phi(\alpha_z)/[1 - \Phi(\alpha_z)\,]$$

$$and\ \delta(\alpha_z) = \lambda(\alpha_z)[\lambda(\alpha_z) - \alpha_z]$$

# Heckman model – Basic Setup (1)

**Selection equation:**

$$z_i^* = \omega_i \gamma + \mu_i$$

$$z_i = \begin{cases} 1 \; if \; z_i^* > 0 \\ 0 \; if \;\; z_i^* \leq 0 \end{cases}$$

**Outcome equation:**

$$y_i = \begin{cases} X_i \beta + \epsilon_i \; if \; z_i^* > 0 \\ - \qquad\qquad if \;\; z_i^* \leq 0 \end{cases}$$

**Assumptions:**

$$\mu_i \sim N(0,1)$$

$$\epsilon_i \sim N(0, \sigma^2)$$

$$corr(\mu_i, \epsilon_i) = \hat{\delta}\rho$$

# Heckman model – Basic Setup (2)

**Conditional Means:**

$$\begin{aligned} E[y_i | y_i \ is \ observed] &= E[y_i | z_i^* > 0] \\ &= E[x_i\beta + \epsilon_i | \omega_i\gamma + \mu_i > 0] \\ &= x_i\beta + E[\epsilon_i | \omega_i\gamma + \mu_i > 0] \\ &= x_i\beta + E[\epsilon_i | \mu_i > -\omega_i\gamma] \end{aligned}$$

where $E[\epsilon_i | \mu_i > -\omega_i\gamma] = \rho\sigma_\epsilon\lambda_i(\alpha_\mu)$

**Outcome Equation:**

$$\begin{aligned} y_i | z_i^* > 0 &= x_i\beta + \rho\sigma_\epsilon\lambda_i(\alpha_\mu) + \upsilon_i \\ &= x_i\beta + \beta_\lambda\lambda_i\left(\alpha_\mu\right) + \upsilon_i \end{aligned}$$

where $\alpha_\mu = \frac{\omega_i\gamma}{\sigma_\mu}$ and $\lambda(\alpha_\mu) = \frac{\phi\left(\frac{\omega_i\gamma}{\sigma_\mu}\right)}{\Phi\left(\frac{\omega_i\gamma}{\sigma_\mu}\right)}$ (Inverse Mills Ratio)

# Heckman’s Two-Step Procedure (1)

## Step 1: Estimate the selection equation

- Use MLE to estimate the Probit equation to obtain estimates of $\gamma$
- For each observation compute $\hat{\lambda}_i = \frac{\phi(\omega_i^{'}\widehat{\gamma})}{\Phi(\omega_i^{'}\widehat{\gamma})}$ and $\hat{\delta}_i = \hat{\lambda}_i(\hat{\lambda}_i + \omega_i^{'}\widehat{\gamma})$

## Step 2: Estimate the outcome equation

- For each observation attach the calculated $\hat{\lambda}_i = \frac{\phi(\omega_i^{'}\widehat{\gamma})}{\Phi(\omega_i^{'}\widehat{\gamma})}$
- Use OLS to estimate $\beta \ and \ \beta_\lambda = \rho\sigma_\epsilon \quad in \quad y_i|z_i^* > 0 = x_i\beta + \beta_\lambda\lambda_i(\alpha_\mu) + \upsilon_i$
- i.e. Estimate $\beta \ and \ \beta_\lambda$ by OLS of y on x and $\hat{\lambda}$

# Heckman's Two-Step Procedure (2)

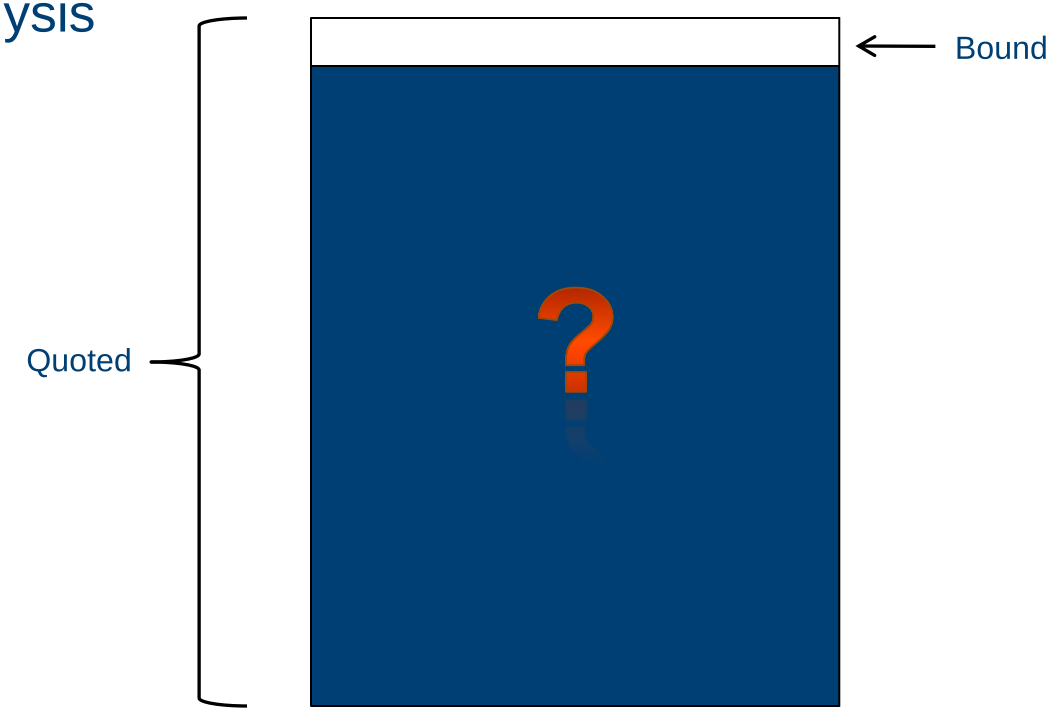
- Assumptions
  - $\mu_i$ and $\varepsilon_i$ are independent of the explanatory variables
  - They both have mean 0
  - $\mu_i \sim N(0,1)$

- Additional notes
  - Non-linearity is introduced via the Inverse Mills ratio
  - The selection and outcome equations do not include the same set of explanatory variables
  - If the selection model does a poor job of determining selection, the outcome equation may provide poor estimates
  - The significance of the coefficient of the Inverse Mills ratio will indicate if there is selection bias

# Some Insurance Applications

- Non-Pricing Applications
  - Commercial Lines: premium audits
  - Homeowners: home inspections
  - Personal Auto: MVRs
  - Competitive analysis

- Pricing



# Parting Comments

- This is a broad topic
- Hundreds of possible statistical methods exist
- Selection bias is present in many insurance processes
- We can improve our analysis by utilizing appropriate techniques to adjust for selection bias

# Short Bibliography

- Amemiya, Takeshi (1984) "Tobit Models: A Survey." Journal of Econometrics 24
- Heckman, J. J. (1976) "The Common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimation for such models." *Annals of Economic and Social Measurement*, 5, (4)
- Heckman, J. J. (1979) "Sample selection bias as a specification error." *Econometrica*, 47 (1)
- Greene, William H. (2008) Econometric Analysis
- Tobin, J. "Estimation of relationships for limited dependent variables." *Econometrica* 26: 24-36
- Vella, F. (1998) "Estimating Models with Sample Selection Bias: A Survey." *Journal of Human Resources*, 33
- Weisberg, Herbert I., PhD. (2010) Bias and Causation: Models and Judgment for Valid Comparisons

# Selection Bias

Thank You!

VALEN
technologies